第26卷 第2期
Vol.26 № 2
计 算 机 工 程
Computer Engineering
2000年2月
February 2000

·人工智能及应用·
文章编号：1000-3428(2000)02-0056-03
中图分类号：TP18；G354.4

本体论与信息检索

廖明宏
(哈尔滨工业大学计算机科学与工程系，哈尔滨150001)

摘 要：试图对本体论做形式化的描述，并在此基础上讨论本体论的一个应用：基于本体论的信息检索方法，它克服了目前基于关键词检索所造成的信息冗余和丢失的不足。其查询方法更符合人类的思维习惯，查询结果也更合理、可用。

关键词：本体论；信息检索；企业知识管理

Ontology and Information Retrieval

Liao Minghong
(Department of Computer Science and Engineering,Haerbin Institute of Technology,Haerbin150001)

【Abstract】This paper tries to describe the ontology in a formal way and introduce one of its application: an information retrieval method based on ontology. This method overcomes the insufficiency of information overloading and missing caused by the keyword-based retrieval methods. So it more conforms to the custom of human thought, and the results become more rational and useful. This method is verified by an experimental protosystem.

【Key words】Ontology; Information retrieval; Enterprise knowledge management

本体论(或称实体论)这个概念在计算机科学中扮演着越来越重要的角色[1]。然而，到目前为止，在计算机界还很难为本体论下一个确切的定义，但斯坦福大学的Gruber给出的定义得到许多同行的认可，即本体论是对概念化的精确描述[2]。本体论的最终目标是精确地表示那些隐含(或不明确的)信息，使得它们可被软件系统重用和共享。

随着因特网的普及与推广，人们越来越依赖于万维网从事科研及商务活动。因而在网上检索信息成了人们最关心的话题之一。常听到有人抱怨，利用现有检索工具来查询某一信息，得到的结果是一堆信息垃圾，很少有他们想要的东西，原因在于目前的信息检索工具主要是基于关键词或内容分类目录进行查找的，很少做进一步的智能化处理。本文借鉴本体论的基本思想，提出一种基于本体论的信息检索方法。这种方法使查询结果更能满足用户的要求。

1 概念化与本体论

本体论是对概念化的精确描述，然而本体论与概念化仍存在本质的区别。在人工智能教科书中，概念化被定义为一个结构<D,R>，其中D表示一个域，R为D上的关系。R是一般数学意义上的关系，故称为外延函数(Extensional relations)。然而，这里所关心的是这些关系本身的含义，称为内涵关系(Intensional relations)或概念关系(Conceptual relations)。如果说外延关系是定义在某个域上的，那么内涵关系则是定义在某个域空间上。

定义1：域空间是一个结构<D，W>，其中D表示一个域，W表示域中元素相关状态的集合(或称可能的世界)。

定义2：n元概念关系：:域空间<D，W>上定义的n元概念关系是一个从W到D上所有n元关系集合的全函数，即：W。

定义3：概念化：域D的概念化是一个三元组C=<D,W,Rc>，其中，Rc表示域空间<D,W>上的概念关系集合。

定义4：语言的内涵解释是一个结构<C,IC>，其中C=<D,W,RC>是一个概念化，而函数 将D中的元素赋上V中的常量符号，RC元素赋上V中的谓词符号。

语言的内涵解释也称为语言的本体论的承诺(Ontological Commitment)。若K=<C,IC>表示语言L的本体论的承诺，我们说L通过K承诺C，而C是K的基本概念化。概念化与本体论的联系与区别在于给定一个带有本体论的承诺K的语言L，L的本体论是一个公理集合，该公理集合使得它的模型集合尽可能近似于L关于K的原有模型。事实上这样的公理集合是不易找到的，因此我们说一个语言L的本体论近似于一个概念化C，如果存在一个本体论的承诺K=<C，IC>，使得L关于K的原有模型被包含于本体论的模型之中。

2 基于本体论的信息检索方法

目前的信息检索方法主要是基于关键词或目录分类的，其查询结果往往产生大量毫无相关的信息，同时又可能丢失重要的信息。由于本体论刻划了事物之间的内在联系，借助于本体论，可以使检索的信息更能满足用户的需求。下面以查询企业雇员能力为例，介绍一种启发式的信息检索方法。

在现代化企业管理中人们越来越清楚地认识到知识，尤其是雇员的知识是企业非常重要的财富。如何组织、管理好这些知识财富成为现代管理重要的研究课题[3,4]。

传统的做法是建立一个数据库系统来管理企业雇员的能力。然而，基于关键词的查询有时是不能满足人们的要求的。比如查找一个懂得"数据库"的雇员来从事一个项目的开发工作，假设在建立的能力查询系统中找不到一个人懂得"数据库"，系统只能返回空记录给用户。而事实上存在某些雇员懂得"关系数据库"或"面向对象数据库"，按常识，这些人应该懂得数据库的基本理论，他们可作为候选人提交给用户。但传统的数据库做不到这一点。当然我们可以通过进一步输入这些关键词达到这一目的，但我们不能保证用户了解"数据库"所有的子领域。而借助于本体论，可以实现这一目

作者简介：廖明宏(1966～)，男，博士，主研方向：人工智能、并行、分布式处理

收稿日期：1999－06－07

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的。假设我们已经为企业建立一个本体论，其中某一部分由图1所示。

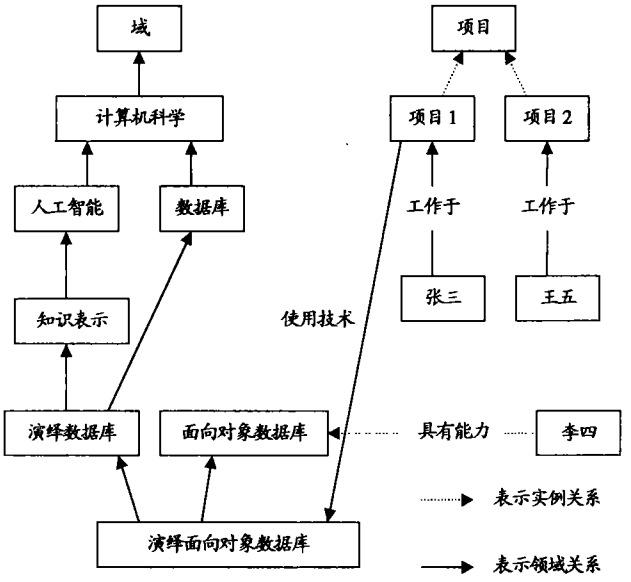

图1 一部分本体论

现在再回到查询懂得数据库的雇员。假设不能得到结果，那么由图1可以看到，李四具有"面向对象数据库"的能力，按常理，他也同样具备"数据库"的能力。一般认为，具备某领域的子领域能力的雇员，同样具备该领域的知识。

我们把问题进一步推广到项目因素。现要查询具备"演绎面向对象数据库"能力的雇员，同样没有结果。然而从图1可以知道张三在干项目1，而项目1应用了"演绎面向对象数据库"技术。因此张三可作为候选人输出。一般地，基于本体论的启发式表达式可定义为n个函数的组合运算：

$$f_1\ f_2\ \ldots f_n \qquad (1)$$

其中，$f_i(\lambda)^\gamma$

λ表示一条有向边e或其反向(写成e^{-1})；

γ表示偏序闭包，它表示路径长度，可取下面任一种值：n，n...m，$\geqslant$n，*(表示$\geqslant$0)或+(表示$\geqslant$1)。

公式(1)说明它以一有向图的结点集作为输入，对每个结点，沿着公式所说明的边以从右到左的顺序进行计算，每步产生的中间结果将作为下一步的输入。

例如上述例子可用启发式公式表示为：

(1) $(\text{has Competence}^{-1})^1$;

(2) $(\text{has Competence}^{-1})^1$ $(\text{is Sub Field OF}^{-1})^+$;

(3) $(\text{work In}^{-1})^1$ $(\text{use Technology}^{-1})^1$

说明：首先沿has Competence边的反向直接查找指向某一概念的人；其次查找指向该概念所有子概念的人。启发式查询算法可描述如下：

算法1：启发式信息检索算法

输入：查询关键词集ks；启发式公式$f_1\ f_2\ \ldots f_n$;

输出：满足启发式条件的所有记录；

方法：

```
begin
(1) 初始化堆栈stack为空;
(2) for i=1 to n do
        push (stack, fi); //fi进栈;
(3) temp=ks;
(4) while stack 非空 do // 从右向左计算fi;
begin
(5)     f=pop(stack);
(6)     temp=f(temp);
      end;
(7)  return temp;
    end;
```

定理1： 算法1的时间复杂性为O(en)，其中n为启发式公式中的函数个数，e为有向图的边的条数。

证明： 算法第(6)步最费时间，由于$f_i=(\lambda)r$，当r为*时(即求所有子树)，它要求遍历一棵子树，若采用先深遍历算法，其时间复杂性为O(e)。算法第(4)步循环n次，故整个算法时间复杂性为O(en)。

3 一个实验原型系统

为验证上述介绍的启发式信息检索思想，我们基于本体论设计并实现了一个企业雇员能力检索系统(CRS)。该系统基于客户/服务器模型，采用Java编程；因此其客户端可在因特网上用Netscape或Internet Explorer浏览器调用，而服务器部分安装在Sun工作站上。CRS的体系结构见图2。

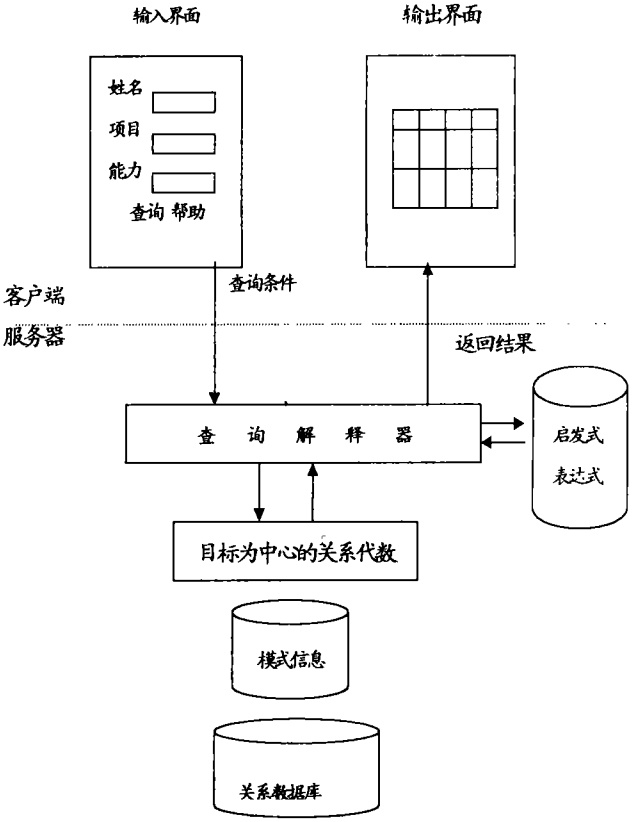

图2 CRS体系结构

在输入部分，用户可查询某一能力，某个人具备的能力或某一项目组成员的能力。其中每个输入域也可以是一个布尔表达式。由客户端提交的查询条件以一定的传输协议送到服务器上。服务器的查询解释根据查询条件和启发式表达式形成关系代数的操作序列，完成相应的操作，并将查询结果

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

以表格方式形成一个HTML文件，回送到客户端。本体论以及其它数据是以关系数据库的形式存放在服务器上的，模式信息描述本体论的结构模式，在它们的基础上设计一套以对象为中心的关系代数(Object-Centered Relational Algebra, OCRA，见文献[5])，它提供一种对当前数据具有演绎能力的面向对象的观点和访问方法。有关本体论的具体设计思想本文不作介绍，有兴趣的读者可查阅文献[6]。实验表明，该系统可充分利用本体论的主要信息，在算法1的思想指导下，有效地检索信息，其查询结果更为合理、实用。

4 结论

本体论及其应用在计算机界得到越来越多学者的重视，然而有关本体论的研究还处在雏形阶段，没有统一的定义和固定的应用领域。本文在澄清概念的基础上，讨论了本体论应用于信息检索的新尝试，并设计了一个实验原型系统。实验结果表明，基于本体论的信息检索方法更符合人类的思维习惯，它可克服传统检索方法造成的信息冗余或信息丢失的缺点，其查询效果更为合理可用。进一步的研究工作将考虑以本体论为核心，设计基于本体论的信息系统。

参考文献

1 Guarino N.Formal Ontology and Information Systems.In Guarino N(ed.):Formal Ontology in Information Systems.Proc.of FOIS'98, Trento,Italy,1998-06-08

2 Gruber T R.Toward Principles for the Design of Ontologies Used for Knowledge Sharing.Int. Journal of Human and Computer Studies,1995:907-928

3 Leary D O.Enterprise Knowledge Management.IEEE Computer, 1998,13(3):54-61

4 Liao M,Hinkelmann K,Abecker A,et al.A Competence Knowledge Base System as Part of the Organizational Memory.in F.Puppe (ed.) XPS99,5th Biannual German Conf.on Knowledge Based System,Springer-verlag,LNAI 1570,Wuerzburg,Germany,1999: 125-137

5 Abecker A,Bernardi A,Hinkelmann K,et al.Techniques for Organizational Memory Information Systems.DFKI Document D-98-02,1998-02

6 Liao M,Abecker A,Bernrdi A,et al.Ontologies for Knowledge Retrieval in Organization Memories.In Workshop on"Learning Software Organization"in SEKE'99,Kaiserslautern,Germany,1999

※※

(上接第31页)

(4)设备及其属性的查询、统计　输入各种设备的名称或设备编码，能在地图上准确定位到该设备，查看其属性信息；对某类设备，根据用户需要统计各种数据，如变压器总容量，线路总长度等，根据需要采用专题地图的方式将数据图形化，使数据以更直观的形式在地图上体现出来。

(5)语音信息接入　调度任务操作完毕后，现场操作人员通过电话按键输入所操作的设备编码及其状态，由语音处理系统得到，通过DDE(动态数据交换)机制，一方面将所得到的信息传输给调度可视化系统，另一方面触发搜索子程序，根据所得到的设备编码及其状态，搜索配电网，执行可视化任务。语音处理系统用C语言开发，具体实现将另文介绍。

(6)调度任务可视化　主要是实现搜索子程序，根据所得到的设备编码及其状态，搜索配电网，用不同的颜色渲染相应线路、设备等，以反映电网实际运行状态。并将服务器中数据实时刷新，使其他终端能显现电网实时运行状况。

此外，操作票的开写是电力调度的一项关键任务。操作票的正确与否影响电力运行的安全。基于本系统，输入一次调度任务所需的操作序列，改变相应的设备状态，可实现操作票的图形验证，有助于操作票的正确开写。

4 结论

基于GIS技术，本文开发了一个配电网调度可视化系统，不仅方便配电网及其设备的信息管理，还可用图形实时显示调度任务执行情况，有助于调度任务决策。本系统已通过福州市电业局验收，安装在调度监控室，运行良好。在此基础上为电业局提供更具体的决策支持，开发基于GIS的决策支持系统，是我们进一步的工作。

致谢: 本文的工作是在福州市配电地理信息系统基础上的进一步研究开发，对该系统工作组所做的基础性工作表示感谢；对福州市电业局陈光捷总工，陈志武副总工，林朝辉，林敏辉等的大力支持和协助表示诚挚谢意!

参考文献

1 吴信才.地理信息系统的基本技术与发展动态.地球科学,1998-04

2 王家耀.地图学与地理信息系统的现状和趋势.测绘通报,1997-06

3 王建华,祝国瑞,毋河海.GIS中可视化媒体的描述和表达.测绘通报,1999-02

※※

(上接第38页)

数据库的选择可以有多种，但也同样要选择安全保护较强的一种，就性能价格比来说，以Microsoft公司的SQL Server 6.5较为合适。

工作站的操作系统可用Windows 9X，CAD工作站选用AuToCAD R13以上版本和VC++，若选用AuToCAD R14，也可选择VB或Delphi作为图纸设计系统的开发工具，本地的数据库可用Access。管理系统软件可采用开发者熟悉的一种，如VB、Delphi、PowerBuilder等。

4 结束语

本系统的基本模型已在常柴股份有限公司实现并运转，由于系统是根据企业的需求和管理模式进行设计的，因此专用性较好，通用性不足。

本系统的运行使企业产品开发部门内部管理更加规范，产品信息的组织和存放更加合理，基本杜绝了因零件修改而带来的混乱，可为企业将来进行更高层次的管理(如PDM、MRPII等)积累管理经验和提供规范化的产品信息数据。

参考文献

1 张奔,王志新.基于ODBC的Visual C++应用程序设计研究.计算机应用,1998,18(11):11-14

2 苟凌怡,魏生民.在Oracle中产品结构树的生成与查询优化.计算机工程与应用,1999,35(7):63-76

3 Microsoft 公司著.希望图书创作室译.Microsoft SQL Server6.5 资源指南.北京:科学出版社,龙门书局,1998

4 康博创作室译. Visual C++ 5 开发使用手册.北京:机械工业出版社,西蒙与舒期特国际出版公司,1998

© 1994-2011 China Academic Journal Electronic Publishing House. All rights reserved.   http://www.cnki.net